Learning Agency Exercise

1. To begin, please review the dataset linked above.

Identify three limitations within the dataset and its documentation. You are encouraged to brainstorm ways to enhance the dataset to make it more valuable for the data science, machine learning, and education research communities. Please limit your response to a maximum of 250 words.

The first limitation of this dataset is its size. While the number of total students is adequate (N=151 is acceptable), there are a few columns with limited information. For example, Imperial College London, Caltech, and University of Cambridge each have only 2 or 1 students accounted for. This means that there will be difficulty predicting how students in these universities would score based on this data alone. Additional classification algorithms could be performed if there were more students in those universities or the other columns in the dataset.

The next limitation is in the documentation and definition. The "Domain" column is said to "indicate whether the exam was divided into two parts". This seems to suggest a binary variable but the column is a categorical variable. This calls into question the understanding of the test definition because a students with a Generic domain has both a general management score and a domain specific score. In addition, it is unclear how rank (and percentile) was given since there is no distinction other than scores and no hierarchy to the other columns provided.

The final limitation is the lack of numerical student data and additional student information that could be gathered to enhance the findings of this data. For example, including student GPA could help make more accurate predictions about final scores rather than mostly categorical data. A survey could also be conducted to ask about a student’s post-grad job status to analyze the effectiveness of this exam on job placement.

2. Analyze a particular trend or pattern within the dataset using your preferred programming language. We encourage you to utilize inferential statistics and be prepared to present your code for review. Additionally, create a slide deck to visually communicate your findings and summarize your results in both of the following formats:

Written: Provide a written summary of your analysis and findings within a 250-word limit.

The first analysis was a general analysis of the data looking at histograms to determine the shape of the data. The scores were mostly left skewed whereas the rank and percentile were fairly uniform. Next, five bar graphs were created grouped by the five independent variables used in the regression analysis (University, Program Name, Specialisation, Domain, and Semester) and their average Percentile. This visualizes the percentile in each variable and if there were any noticeable differences but based on these bar graphs, there were none. This again, may be the result of a small sample size in each group attributing to large standard deviations. In the case of Imperial College, there was only 1 data point so there was no standard deviation.

Then regression analysis conducted was to determine a linear model that would predict percentile based on the dependent variables. As part of the preparation of the data, the Semester column was converted into numerical data and training and testing datasets were created. The categorical data were dummy coded using the scikit-learn Ordinal Encoder and then standardized using the scikit-learn Standard Scaler. A linear regression model was created with an RMSE of almost 0 and an R-squared value of 1. This seems to indicate overfitting. Thus a stochastic gradient descent regressor was also conducted which yielded an RMSE of about .002 and R-squared of 0.968. This value seems to be more reasonable and a plot of the test vs. predicted values were also created.

Spoken: Create a video or audio recording (maximum 5 minutes) in which you verbally present your slide deck.